Sequence capture and next-generation resequencing of multiple tagged nucleic acid samples for mutation screening of urea cycle disorders

Amstutz, U ; Andrey-Zürcher, G ; Suciu, D ; Jaggi, R ; Häberle, J ; Largiadèr, C R

Abstract: BACKGROUND: Molecular genetic testing is commonly used to confirm clinical diagnoses of inherited urea cycle disorders (UCDs); however, conventional mutation screenings encompassing only the coding regions of genes may not detect disease-causing mutations occurring in regulatory elements and introns. Microarray-based target enrichment and next-generation sequencing now allow more-comprehensive genetic screening. We applied this approach to UCDs and combined it with the use of DNA bar codes for more cost-effective, parallel analyses of multiple samples. METHODS: We used sectored 2240-feature medium-density oligonucleotide arrays to capture and enrich a 199-kb genomic target encompassing the complete genomic regions of 3 urea cycle genes, OTC (ornithine carbamoyltransferase), CPS1 (carbamoyl-phosphate synthetase 1, mitochondrial), and NAGS (N-acetylglutamate synthase). We used the Genome Sequencer FLX System (454 Life Sciences) to jointly analyze 4 samples individually tagged with a 6-bp DNA bar code and compared the results with those for an individually sequenced sample. RESULTS: Using a low tiling density of only 1 probe per 91 bp, we obtained strong enrichment of the targeted loci to achieve 90% coverage with up to 64% of the sequences covered at a sequencing depth 10-fold. We observed a very homogeneous sequence representation of the bar-coded samples, which yielded a >30% increase in the sequence data generated per sample, compared with an individually processed sample. Heterozygous and homozygous disease-associated mutations were correctly detected in all samples. CONCLUSIONS: The use of DNA bar codes and the use of sectored oligonucleotide arrays for target enrichment enable parallel, large-scale analysis of complete genomic regions for multiple genes of a disease pathway and for multiple samples simultaneously. This approach thus may provide an efficient tool for comprehensive diagnostic screening of mutations.

DOI: https://doi.org/10.1373/clinchem.2010.150706

Clinical Chemistry 57:1
102–111 (2011)

Sequence Capture and Next-Generation Resequencing of Multiple Tagged Nucleic Acid Samples for Mutation Screening of Urea Cycle Disorders

Ursula Amstutz,[1,2] Gisela Andrey-Zürcher,[1] Dominic Suciu,[3] Rolf Jaggi,[4] Johannes Häberle,[5] and Carlo R. Largiadèr[1*]

BACKGROUND: Molecular genetic testing is commonly used to confirm clinical diagnoses of inherited urea cycle disorders (UCDs); however, conventional mutation screenings encompassing only the coding regions of genes may not detect disease-causing mutations occurring in regulatory elements and introns. Microarray-based target enrichment and next-generation sequencing now allow more-comprehensive genetic screening. We applied this approach to UCDs and combined it with the use of DNA bar codes for more cost-effective, parallel analyses of multiple samples.

METHODS: We used sectored 2240-feature medium-density oligonucleotide arrays to capture and enrich a 199-kb genomic target encompassing the complete genomic regions of 3 urea cycle genes, *OTC* (ornithine carbamoyltransferase), *CPS1* (carbamoyl-phosphate synthetase 1, mitochondrial), and *NAGS* (*N*-acetylglutamate synthase). We used the Genome Sequencer FLX System (454 Life Sciences) to jointly analyze 4 samples individually tagged with a 6-bp DNA bar code and compared the results with those for an individually sequenced sample.

RESULTS: Using a low tiling density of only 1 probe per 91 bp, we obtained strong enrichment of the targeted loci to achieve ≥90% coverage with up to 64% of the sequences covered at a sequencing depth ≥10-fold. We observed a very homogeneous sequence representation of the bar-coded samples, which yielded a >30% increase in the sequence data generated per sample, compared with an individually processed sample. Heterozygous and homozygous disease-associated mutations were correctly detected in all samples.

CONCLUSIONS: The use of DNA bar codes and the use of sectored oligonucleotide arrays for target enrichment enable parallel, large-scale analysis of complete genomic regions for multiple genes of a disease pathway and for multiple samples simultaneously. This approach thus may provide an efficient tool for comprehensive diagnostic screening of mutations.

Urea cycle disorders (UCDs)[6] are inborn errors of metabolism caused by a reduced or absent activity of enzymes involved the transfer of nitrogen from ammonia to urea *(1)*. Because the urea cycle is the only pathway capable of metabolizing excess nitrogen, such enzyme defects lead to toxic hyperammonemia, which usually has a high morbidity and mortality *(2)*. The urea cycle consists of 6 nuclear genome–encoded enzymes, 3 of which are located in the mitochondrial matrix: *CPS1*[7] (carbamoyl-phosphate synthetase 1, mitochondrial), *NAGS* (*N*-acetylglutamate synthase), and *OTC* (ornithine carbamoyltransferase). With the aid of the allosteric activator *N*-acetylglutamate synthesized by NAGS, CPS1 catalyzes the transformation of ammonia to carbamoyl phosphate. OTC subsequently metabolizes carbamoyl phosphate to citrulline, which is then converted to urea in the cytosol.

Whereas measurements of intermediary metabolites are primarily used to establish diagnoses of inborn UCDs, genetic mutation testing is commonly used to confirm diagnoses. Today, the standard approach to mutation screening for OTC, CPS1, or NAGS deficiency encompasses the sequencing of coding regions and splice sites; however, this approach detects a disease-causing mutation in only about 80% of pa-

[1] Institute of Clinical Chemistry, Inselspital, University Hospital and University of Bern, Bern, Switzerland; [2] Pharmaceutical Outcomes Programme, Child & Family Research Institute, University of British Columbia, Vancouver, British Columbia, Canada; [3] CustomArray, Inc., Mukilteo, WA; [4] Department of Clinical Research, University of Bern, Bern, Switzerland; [5] Division of Metabolism, University Children's Hospital, Zurich, Switzerland.
* Address correspondence to this author at: Institute of Clinical Chemistry, Inselspital, IKC INO F, CH-3010 Bern, Switzerland. Fax +41-31-632-48-62; e-mail carlo.largiader@insel.ch.
Received May 21, 2010; accepted October 19, 2010.
Previously published online at DOI: 10.1373/clinchem.2010.150706
[6] Nonstandard abbreviations: UCD, urea cycle disorder; NGS, next-generation sequencing; MSC, microarray-based sequence capture; T_m, melting temperature.
[7] Human genes: *CPS1*, carbamoyl-phosphate synthetase 1, mitochondrial; *NAGS*, *N*-acetylglutamate synthase; *OTC*, ornithine carbamoyltransferase.

tients with OTC deficiency *(3)*. This fact, together with the recent identification of deleterious mutations located deep within the introns of the *OTC* gene *(4, 5)*, demonstrates the need for new tools for mutation screening that include investigation of the noncoding regions of genes. Furthermore, for very large genes such as *CPS1* (>120 kb encompassing 38 exons), being able to conduct a comprehensive analysis at the genomic DNA level without the need for numerous PCR amplifications of single exons could greatly simplify the search for disease-causing mutations. Finally, given that NAGS supplies the cofactor required for the CPS1-catalyzed reaction, biochemical measurements of intermediary metabolites cannot be used to distinguish between NAGS and CPS1 deficiencies *(2)*. Similarly, assays of biochemical markers can yield ambiguous results in late-onset forms of OTC deficiency *(6–8)*. In such cases, the ability to simultaneously screen multiple genes potentially involved in a suspected metabolic disorder could avoid invasive procedures required for direct measurements of enzyme activity *(9, 10)* and thereby improve diagnostic efficiency.

With the development of next-generation sequencing (NGS) technologies, it is now possible to generate large amounts of sequence data at lower cost and with less effort *(11)*, offering new possibilities for diagnostic mutation screening *(12–16)*. To analyze specific regions in complex eukaryotic genomes with NGS, however, requires prior selection or amplification of the target regions. PCR, the most widely applied method for this purpose, has strong limitations in terms of multiplexing grade and product length that make it unsuitable for large or dispersed genomic targets. Therefore, hybridization to complementary oligonucleotide probes, either on DNA microarrays *(17–19)* or in solution *(20)*, has recently been successfully used to enrich genomic regions for NGS.

Given that the amount of sequence data that most NGS instruments generate in a full sequencing run is much larger than the amount required for reliable variant detection in moderately sized (200 kb) targets, analyzing multiple samples in parallel via the use of DNA bar codes might allow optimization of the cost-effectiveness of this approach *(13, 21)*. The addition of a specific sequence tag (DNA bar code) to each sample enables pooled processing of samples (thereby reducing time and materials), as opposed to physical separation (e.g., sectoring of sequencing plates), which is also associated with a reduction in the total number of sequences generated *(13, 21)*.

Most microarray-based sequence capture (MSC) protocols have used high-density oligonucleotide arrays containing up to 385 000 capture probes to enrich genomic targets *(18, 22)*. We assessed the suitability of the use of medium-density arrays containing only 2240 oligonucleotide probes combined with a long-read NGS technology for MSC and the use of DNA bar codes for the parallel analysis of multiple samples. With inherited UCDs as a model, we investigated a 199-kb region that encompasses the complete genomic sequences of *OTC*, *NAGS*, and *CPS1*. This approach enabled (*a*) the detection of mutations not only in coding regions but also in regulatory and intronic regions, (*b*) the comprehensive investigation at the genomic DNA level of a large gene comprising many exons; and (*c*) the simultaneous investigation of 3 genes involved in the same disease pathway.

Materials and Methods

DNA SAMPLES

We analyzed a total of 4 DNA samples—2 samples taken from carriers of disease-causing *OTC* mutations (samples OTCA and OTCB) and 2 samples from CPS1-deficient patients (samples CPSA and CPSB). In all patients, the disease-associated mutations had previously been detected with Sanger sequencing. The patients or their parents provided informed consent for the genetic studies in this initial analysis. Subsequently, samples were irreversibly anonymized, and mutations were not known to the investigator. As a negative control, we also sequenced 1 sample without performing the MSC hybridization (sample UNC). We used DNA bar codes to analyze samples CPSA, CPSB, OTCB, and UNC, whereas we analyzed sample OTCA without bar codes by use of a sectored 454 picotiter plate in a comparison of both approaches.

ARRAY DESIGN AND SYNTHESIS

The target region encompassed the entire genomic regions of *OTC*, *NAGS*, and *CPS1*, along with 6 control loci for the assessment of enrichment efficacy (see Methods in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol57/issue1). Probes were designed by tiling the target regions with probes of 50–60 bp, which were selected on the basis of melting temperature (T_m), complexity, secondary structure, GC content, and specificity. We ultimately used a total of 2240 capture probes resulting in a mean tiling density of 1 probe per 91 bp. There was no probe redundancy in the final array design. The Methods section in the online Data Supplement provides a more detailed description of the probe design. Arrays containing 4 independent 2240-feature microarrays on a single slide were synthesized on a CustomArray Synthesizer (CombiMatrix).

SAMPLE PREPARATION AND MSC HYBRIDIZATION

The Methods section in the online Data Supplement provides a detailed description of sample preparation and array hybridization procedures. In brief, 20 μg of whole genome–amplified DNA was fragmented by nebulization to fragments of approximately 500–600 bp, blunt-ended, and ligated to universal adaptors. The adaptor-ligated DNA was hybridized for 64 h at 42 °C to one of 4 independent array sectors for samples CPSA, CPSB, and OTCB with hybridization conditions that were slightly modified compared with those for sample OTCA (see Methods in the online Data Supplement). We washed the hybridized arrays, eluted the captured DNA with molecular-grade water at 95 °C, and amplified the DNA with the universal adaptors as primers. For DNA bar coding of samples CPSA, CPSB, OTCB, and UNC, we used primers containing a sample-specific 6-bp sequence tag *(21)* at the 5′ end. We assessed enrichment success by using real-time quantitative PCR of the control loci and comparing samples taken before and after hybridization (see Methods in the online Data Supplement).

454 SEQUENCING AND DATA ANALYSIS

A 454-sequencing service provider (Microsynth) prepared a single sequencing library from the tagged and equally pooled samples CPSA, CPSB, OTCB, and UNC, and a separate library for sample OTCA. The libraries were analyzed on a 454 Genome Sequencer FLX Titanium (Roche) with half a picotiter plate for the pooled library and one-eighth of a plate for sample OTCA. We sorted sequence reads by bar code as described in the online Data Supplement and mapped them to the human genome and target regions with GS Reference Mapper (Roche Applied Sciences) and MOSAIK assembler software *(23)*. Only reads with a unique match to a chromosome containing a target gene were used for mapping to the target sequence. Statistical analyses were performed with the statistics software R *(24)*. A detailed description of these analyses is given in the Methods in the online Data Supplement.

Results

ENRICHMENT AND 454 SEQUENCING OF CAPTURED SAMPLES

The degree of enrichment for the captured samples was estimated by quantitative PCR analysis to be between 1489-fold and 3023-fold, indicating that MSC hybridization was successful (Table 1). For the pooled library of samples CPSA, CPSB, OTCB, and UNC, we obtained a total of 584 390 sequence reads and 181 Mb of sequence with a mean read length of 311 bp. Of these sequence reads, 524 378 (90%) could be assigned unambiguously to one of the samples with aid of the DNA

Table 1. Mapping results, enrichment, coverage, and sequencing depth.

Sample	Reads (unique reads), n[c]	Reads aligned, n[d]	Base pairs aligned, n[e]	-Fold enrichment (qPCR[f])	Coverage, n (%)[a]				Sequencing depth, -fold[b]			
					All (199 465 bp)	*OTC* (70 171 bp)	*NAGS* (5602 bp)	*CPS1* (123 692 bp)	All	*OTC*	*NAGS*	*CPS1*
CPSA	146 115 (131 474)	10 174	3 166 270	1517	186 715/105 127 (93.6/52.7)	62 505/25 672 (89.1/36.6)	5085/689 (90.8/12.3)	119 125/78 766 (96.3/64.9)	10 (4–22)	7 (2–13)	3 (1–5)	14 (6–29)
CPSB	122 264 (109 828)	9225	2 835 704	2102	186 128/103 037 (93.3/51.7)	64 744/35 946 (92.3/51.2)	5294/1302 (94.5/23.2)	116 090/65 789 (93.9/54.8)	10 (4–20)	10 (3–20)	5 (3–9)	11 (4–20)
OTCB	119 664 (108 442)	11 478	3 680 654	3023	189 346/128 355 (94.9/64.3)	65 231/42 733 (93.0/60.9)	4563/958 (81.5/17.1)	119 552/84 664 (96.7/69.6)	14 (6–26)	13 (5–24)	3 (1–7)	15 (7–28)
UNC	136 335 (122 829)	202	64 707	—	31 951/863 (16.0/0.4)	10 553/0 (15.0/0)	0/0 (0/0)	21 398/863 (17.3/0.7)	0 (0–0)	0 (0–0)	0 (0–0)	0 (0–0)
OTCA	91 688 (49 481)	12 712	3 083 113	1489	179 270/93 083 (89.9/46.7)	64 549/37 234 (92.0/53.1)	5298/1018 (94.6/18.2)	109 423/54 831 (88.5/44.3)	8 (3–22)	11 (3–27)	3 (1–7)	8 (2–20)

[a] Data are presented as the numbers of base pairs in the target region covered by ≥1 sequence read and by ≥10 sequence reads, respectively. Corresponding percentages are in parentheses.
[b] Data are presented as the median (interquartile range).
[c] Data are presented as the number of reads obtained (number of reads with a unique match in the human genome).
[d] Number of aligned sequence reads mapping with 95% sequence identity to the target region.
[e] Number of base pairs in aligned sequence reads.
[f] qPCR, real-time quantitative PCR.

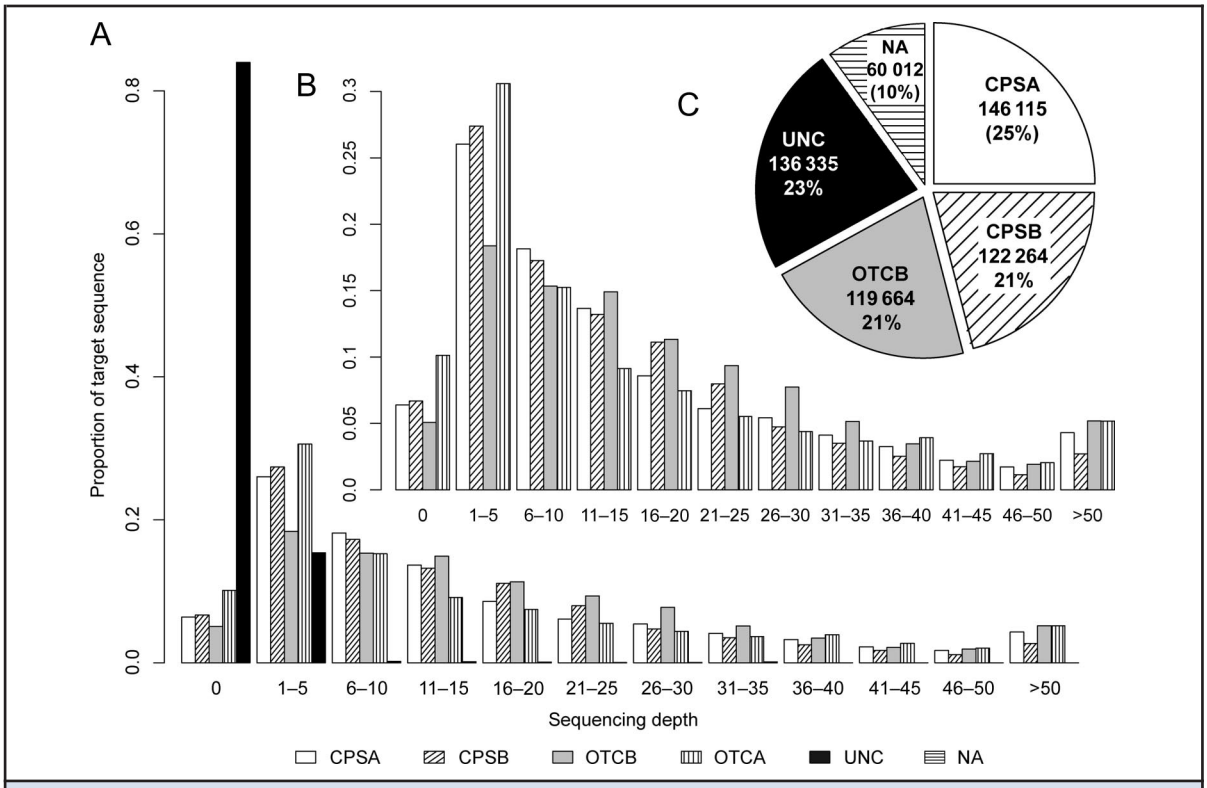

Fig. 1. Sequencing depth and sequence representation.

(A), Distribution of sequencing depth for the 5 analyzed samples. (B), Detailed view excluding sample UNC. (C), Sequence representation of the 4 DNA–bar-coded and pooled samples obtained from a single sequencing library. The pie chart has the same CPSA, CPSB, OTCB, and UNC data as (A) and indicates the number of reads for each bar-coded sample. NA, nonassigned sequences.

bar codes. The remaining sequence reads were either not assigned to any sample or assigned to multiple samples because of sequencing or synthesis errors in the bar code; these reads were therefore removed from the analysis. All 4 samples were represented with very similar proportions of sequence reads (range, 21%–25%), with at least 119 664 reads obtained per sample (Fig. 1B).

For the individually sequenced sample OTCA, we obtained 91 688 sequence reads (20 Mb) with a mean read length of 219 bp. For the samples from the pooled library, 89% of all sequence reads had a unique match in the human genome, whereas this proportion was only 58% for sample OTCA (Table 1 and online Table 1). Of all reads with a unique match to a target chromosome, between 9225 and 12 712 reads (corresponding to >2.8 Mb of sequence per sample) were mapped to the target region with ≥95% sequence identity (Table 1) in the MSC-enriched samples, whereas only 202 reads (64 707 bp) were mapped in sample UNC.

In the MSC-enriched samples, ≥90% of the target region was covered by at least 1 sequence read at a median sequencing depth of 8- to 14-fold (Table 1, Fig. 1A). Sequence coverage and median sequencing depth were lowest in sample OTCA, despite this sample having a similar overall number of mapped base pairs (Table 1, Fig. 1). This result indicated a more homogeneous sequence distribution across the target region for samples CPSA, CPSB, and OTCB, potentially because of the slightly modified hybridization protocol (Table 1). Sample OTCB had the highest sequence coverage and sequencing depth and also had the highest estimated enrichment (Table 1). In this sample, 79.8% of the target sequence had a sequencing depth of ≥5-fold, and 64.3% was sequenced at a depth of ≥10-fold (Table 2).

INTERSAMPLE REPRODUCIBILITY

The 3 samples captured with an identical protocol (CPSA, CPSB, and OTCB) had very similar distributions of sequencing depth, which demonstrated the

Table 2. Intersample reproducibility of sequence coverage.

Depth, -fold	Base pairs of target region sequenced at given depth, n (%)					Overlap[a]			
	CPSA	CPSB	OTCB	OTCA	UNC	3 Samples, n (%)	Percentage of minimum[b]	4 Samples, n (%)	Percentage of minimum[b]
>0	186 715 (93.6)	186 128 (93.3)	189 346 (94.9)	179 270 (89.9)	31 951 (16)	176 724 (88.6)	94.9	165 555 (83)	92.3
≥5	143 264 (71.8)	141 102 (70.7)	159 085 (79.8)	126 837 (63.6)	1339 (0.7)	125 399 (62.9)	88.9	102 988 (51.6)	81.2
≥10	105 127 (52.7)	103 037 (51.7)	128 355 (64.3)	93 083 (46.7)	863 (0.4)	85 090 (42.7)	82.6	65 575 (32.9)	70.4
≥20	57 024 (28.6)	51 971 (26.1)	73 954 (37.1)	57 539 (28.8)	428 (0.2)	35 223 (17.7)	67.8	24 732 (12.4)	47.6

[a] Overlap of sequence covered at the same minimum depth for the 3 samples captured with an identical protocol (CPSA, CPSB, OTCB) and for all 4 captured samples.
[b] Overlap as the percentage of the minimum number of base pairs sequenced at a given depth across the investigated samples.

high intersample reproducibility of the MSC procedure (Table 2, Fig. 2). The 3 samples had correlation coefficients for sequencing depth that ranged from 0.84 to 0.86 (see Table 2 in the online Data Supplement), and 82.6% of the nucleotide positions in the target sequence with a sequencing depth ≥10-fold in the least-enriched sample (CPSB) were sequenced at the same depth or greater in the other samples (Table 2).

DIFFERENCES IN ENRICHMENT BETWEEN THE 3 TARGET GENES

The target genes had similar sequence coverages, but *NAGS* had the lowest median sequencing depth, indicating a less successful enrichment for this gene (Table 1). Comparison of the capture probes for the target genes revealed that *NAGS* probes had, on average, a higher GC content, a shorter length, a higher T_m, and a lower tiling density than probes for *OTC* or *CPS1* (see Table 3 in the online Data Supplement), suggesting that *NAGS* has sequence properties that are more challenging for MSC and NGS.

Furthermore, we observed that *OTC* had a lower median sequencing depth than *CPS1* in sample CPSA. Because sample CPSA was the only sample from a male patient, the lower sequencing depth in the *OTC* gene can be explained by the patient's hemizygosity for this X-linked gene.

NONSPECIFIC SEQUENCE MAPPING

The amount of sequence mapped to the target regions was inversely correlated with the proximity to a capture probe in the MSC-enriched samples; however, no such correlation was observed in sample UNC, a finding that demonstrates the specificity of the MSC approach (Fig. 3A). In the MSC-enriched samples, >91% of the mapped sequence data aligned on or within 100 bp of a capture probe, whereas in sample UNC, a large proportion (42%) of the aligned sequences mapped to regions with no capture probe within 100 bp (Fig. 3A). A majority (94%) of these sequences mapped to a specific section of approximately 1800 bp in *CPS1*, a region for which no capture probes had been designed (Fig. 3B). We had rejected all probes tested for this region because of substantial similarities with multiple other regions in the human genome (Fig. 3B). This finding suggests that the sequences aligning to this section were not captured by MSC but were wrongly aligned from other genomic regions because of the high sequence similarity.

MUTATION DETECTION IN MSC-ENRICHED TARGET REGIONS

All *OTC*, *NAGS*, and *CPS1* exons with flanking intronic and untranslated regions were screened for potential sequence variants, and parts of these screened regions were resequenced with Sanger technology (see Table 4 in the online Data Supplement). With both technologies, we sequenced a total of approximately 34 kb, of

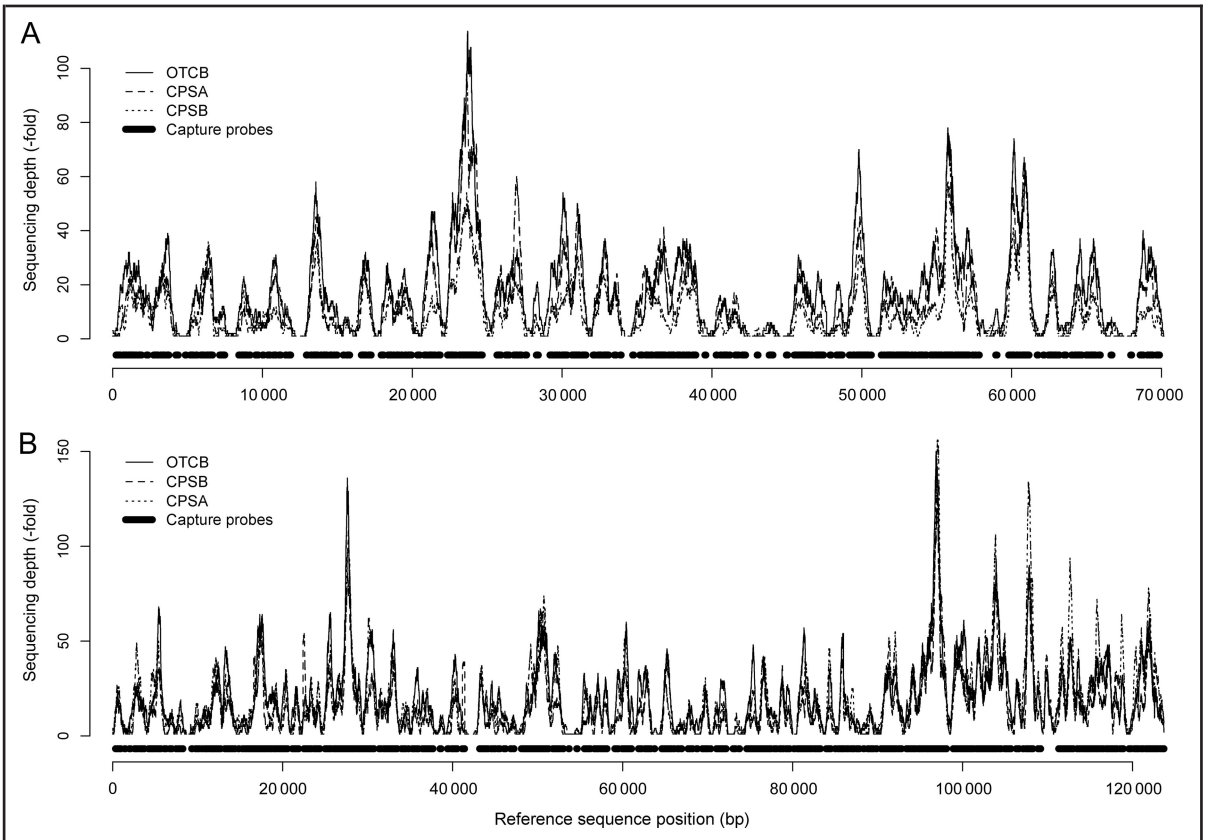

Fig. 2. Reproducibility of the MSC procedure.

(A), Reproducibility of sequencing depth in the *OTC* gene for 3 captured, tagged, and simultaneously analyzed samples. (B), Reproducibility of sequencing depth in the *CPS1* gene for the same 3 samples.

which 20 kb was covered ≥10-fold in the NGS analysis and 27 kb was sequenced at a coverage of ≥5-fold (Table 3). We correctly identified the previously detected disease-associated mutations in all 4 samples, including 4 heterozygous substitutions, 1 homozygous substitution, and 1 heterozygous 3-bp deletion (see Table 5 in the online Data Supplement). For heterozygote calling, we assessed 2 different cutoffs for the proportion of variant alleles observed in the NGS data (Table 3). For both cutoffs, most (>87%) of the detected sequence variants in regions with a sequencing depth ≥10-fold were confirmed with Sanger technology, with the more stringent cutoff (cutoff B) producing an increased positive predictive value but a reduced sensitivity (Table 3). Overall accuracy tended to be increased when the sequencing-depth threshold was increased, although the total number of variants assessed at higher depths was too small for reliable statistical inference. In regions with a sequencing depth between 5-fold and 9-fold, we observed a higher error rate for heterozygous, but not for homozygous, variants (Table 3). Finally, 57 (92%) of 62 positions deviating from the reference sequence at a depth of ≥10-fold, which were not analyzed with Sanger sequencing, had a dbSNP record *(25)*, a result again indicating a low false-positive rate for the NGS data. Insertions or deletions in or close to homopolymer regions of ≥4 repeats of the same base, a threshold selected by visual inspection of the sequencing data, were not considered in all analyses. In such regions, we confirmed only 2 of 10 putative variants with Sanger sequencing, indicating a high false-positive rate.

Discussion

In this study, we successfully used medium-density oligonucleotide arrays for MSC and NGS and used sectored arrays and DNA bar coding to analyze multiple samples in parallel. Capturing the complete genomic regions of *OTC*, *CPS1*, and *NAGS* permitted the simultaneous and comprehensive screening for disease-causing mutations in 3 genes involved in inborn UCDs.

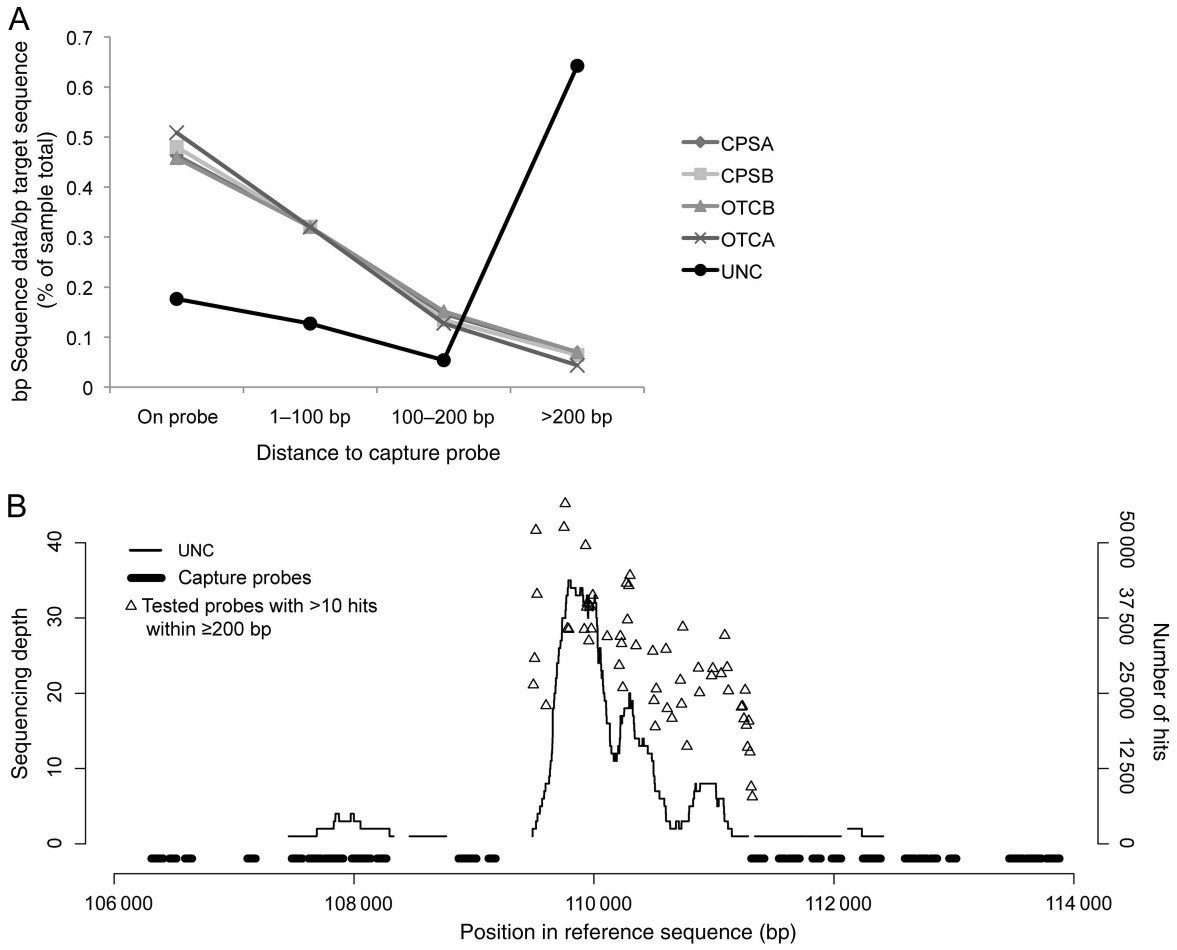

Fig. 3. Specificity of the MSC procedure and nonspecific sequence mapping.

(A), Correlation of the sequence data obtained with the distance to the capture probe. (B), Nonspecific sequence mapping in a 1800-bp region in *CPS1* and similarities of the tested capture probes for this region to sequences in the human genome (consecutive probes with >10 C-BLAST hits spanning ≥200 bp are shown).

MUTATION DETECTION IN MSC-ENRICHED TARGET REGIONS

We correctly identified all disease-causing mutations in the target genes in all samples. Furthermore, Sanger sequencing confirmed the majority of additional polymorphisms detected in the screened regions, a finding demonstrating the potential of this approach for mutation detection, given a sufficient sequencing depth. We obtained a high accuracy with a relatively low threshold depth (≥10-fold) for variant detection. We also reliably detected homozygous variants at a depth of 5-fold to 9-fold, indicating that a lower threshold could potentially be used for hemizygous regions, such as X-linked genes in male patients. As demonstrated by the different cutoffs assessed for calling heterozygous variants, the threshold proportion of variant alleles also strongly influenced the sensitivity and specificity of mutation detection. A higher number of false positives was observed with a threshold of 20% variant alleles, whereas increasing this threshold to 30% produced a loss in sensitivity due to 1 undetected heterozygous mutation. Altogether, we observed only 1 false positive and 1 false negative in 19 615 analyzed base pairs at a 30% threshold and a sequencing depth of ≥10-fold (for a more detailed discussion, see the Supplementary Materials in the online Data Supplement). Diagnostic applications, however, require a higher threshold of sequencing depth to increase the accuracy of heterozygote calling *(16)*.

Further refinement of standard thresholds and reliable estimation of error rates for mutation discovery with NGS requires the validation of larger numbers of putative variants with Sanger sequencing *(14)*. In the

Table 3. Results for the use of MSC and NGS for detecting a variant position deviating from the reference sequence, compared with Sanger sequencing results.

Variant[a]	Deviant allele, %[b]	TP,[c] n	FP, n	FN, n	TN, n	Sensitivity	Specificity	PPV	NPV
Depth ≥20-fold									
Homozygous	≥80	3	0	0	11 727	1	1	1	1
Heterozygous, cutoff A	20–80	11	1	0	11 718	1	>0.99	0.92	1
Heterozygous, cutoff B	30–70	11	0	0	11 719	1	1	1	1
Depth ≥15-fold									
Homozygous	≥80	6	0	0	15 617	1	1	1	1
Heterozygous, cutoff A	20–80	15	3	0	15 605	1	>0.99	0.83	1
Heterozygous, cutoff B	30–70	15	1	0	15 607	1	>0.99	0.94	1
Depth ≥10-fold									
Homozygous	≥80	8	0	0	19 607	1	1	1	1
Heterozygous, cutoff A	20–80	21	3	0	19 590	1	>0.99	0.88	1
Heterozygous, cutoff B	30–70	20	1	1	19 593	0.95	>0.99	0.95	>0.99
Depth ≥5-fold									
Homozygous	≥80	16	0	0	26 646	1	1	1	1
Heterozygous, cutoff A	20–80	23	8	2	26 629	0.92	>0.99	0.74	>0.99
Heterozygous, cutoff B	30–70	22	4	3	26 633	0.88	>0.99	0.85	>0.99

[a] The different proportions of deviant alleles (i.e., cutoff A or B) used for heterozygote calling; given as percentages in column 2.
[b] Percentage of deviant alleles observed in the NGS data used for genotype calling.
[c] TP, true positive (variant confirmed by Sanger sequencing); FP, false positive (variant not confirmed by Sanger sequencing); FN, false negative (variant not detected in NGS analysis but observed in Sanger sequencing); TN, true negative (no deviation from reference sequence with both sequencing methods); PPV, positive predictive value; NPV, negative predictive value.

clinical setting, confirmation of putative disease-associated mutations with Sanger sequencing is thus strongly advisable until these standards become established. In this context, consideration also needs to be given to the increased overall number of sequence variants that will be detected when larger genomic regions are investigated *(13)*. Filtering strategies that identify known polymorphisms and predict potential functional effects are required to distinguish between potential disease-causing mutations and benign variants *(26)*.

DNA BAR CODING

We observed no reduction in sequence quality or increase in error rate due to the pooling of tagged samples, as indicated by the longer read lengths obtained for the bar-coded samples and the very low estimates of error rate for all samples. In contrast, compared with the individually analyzed sample OTCA, use of DNA bar codes yielded a >30% increase in the number of sequence reads per sample, which reduced the cost per nucleotide sequenced. Furthermore, only a single sequencing library needed to be prepared for 4 samples. Having only a single library preparation substantially reduced the costs for library preparation and thus significantly improved the cost-effectiveness of the MSC/NGS approach. In the future, with the continuing increase in read lengths of NGS technologies and the upcoming third-generation instruments promising even longer sequence reads *(27)*, parallel analyses of larger numbers of samples will become possible, because each sample will require fewer reads for sufficient coverage.

REDUCED TILING DENSITY AND THE ADVANTAGE OF LONG READ LENGTHS

Our use in this study of a single MSC hybridization to an oligonucleotide array containing only 2240 capture probes produced enrichment efficacies similar to those obtained with protocols that used >100-fold more probes or that used 2 successive MSC hybridizations for a similarly sized target *(12, 18, 22, 28, 29)*. Our results have thus demonstrated the suitability of this approach for capturing moderately sized targets (approximately 200 kb). Despite the low tiling density, our protocol also yielded enrichment uniformities similar to those reported for other studies *(12, 18, 22, 28, 29)*.

Further studies are required, however, to determine the upper limit of MSC target size with such a low probe number. This successful enrichment with a much lower tiling density may be explained by the higher binding capacity of capture probes to the different array surface we used, compared with other oligonucleotide arrays (see discussion in the Supplementary Materials in the online Data Supplement).

Besides platform-specific issues, which require experimental validation, reduced tiling densities may represent a platform-independent approach to efficiently capture long templates for NGS technologies that generate long sequence reads. Because of their lower diffusion rates and increased intramolecular base pairing, the hybridization of long DNA fragments to capture probes is expected to be inferior to that of short fragments *(30)*. Use of a high tiling density for target enrichment may therefore produce a general bias toward shorter captured fragments, reducing the overall sequence read length. Although a more controlled DNA-fragmentation process may overcome this bias in part, one can expect that a smaller number of short fragments will overlap with capture probes at a lower tiling density. This feature has the potential to reduce this bias and enable more longer fragments to be captured and sequenced. Comparative investigations with other target-enrichment systems are needed, however, to address this issue in more detail.

The possibility of obtaining long sequence reads may be of particular advantage for diagnostic applications. In addition to the less computationally intensive alignment procedure and a higher sensitivity for detecting larger insertions or deletions *(12)*, the use of a long-read NGS technology improves the direct observation of haplotype structures, because each NGS read displays the gametic phase directly. Longer read lengths produce longer stretches of known gametic phase, simplifying haplotype inference and thus providing valuable information for diagnostic applications *(31)*.

TARGET SEQUENCE PROPERTIES AND NONSPECIFIC MAPPING

Despite the strong overall enrichment we observed in this study, enrichment efficacy varied, depending on the genomic target. We observed only moderate enrichment for the *NAGS* gene. In agreement with other studies [e.g., *(16, 20, 31)*], this finding indicates that MSC is more challenging in regions with unfavorable target sequence properties (e.g., high GC content). Furthermore, the amplification bias for the same regions during the whole genome–amplification process used for this protocol may have exacerbated this effect. At present, however, this amplification step may be unavoidable in the clinical setting, where often only small amounts of patient DNA are obtainable. Further optimization of the criteria used for probe design may improve enrichment in such regions, however. Moreover, including multiple probe replicates on an array or locally increasing the tiling density for problematic regions may compensate for a poorer enrichment or potential amplification bias and thus improve enrichment. Finally, further refinement of the enrichment procedure may also improve capture efficiency in such regions *(32)*.

Interestingly, a substantial amount of sequences in the uncaptured sample UNC mapped to a segment in the target sequence that contained no capture probes because of high sequence similarities to multiple other regions in the genome. This finding is surprising because only sequence reads with a unique match to the target chromosomes were used in the analysis, and it indicates that genomic regions with many highly homologous sequences present a challenge for NGS mapping algorithms. Because the sequencing depth in sample UNC in this region was above the threshold used for mutation detection, such nonspecific mapping represents a potential source of false positives in a mutation screening *(12)*. Further refinement of mapping algorithms is required to permit discrimination between such highly similar sequences *(12)*.

Conclusions

Our results highlight the potential of MSC combined with NGS for diagnostic mutation screening, as we have demonstrated for inherited UCDs. The use of sectored arrays and the use of DNA bar codes are independent possibilities for increasing throughput and reducing analysis costs. This approach not only enables the simultaneous and comprehensive investigation of multiple target genes but also facilitates the analysis of multiple samples in parallel, greatly simplifying the search for disease-causing mutations. Finally, the ability to investigate complete coding, intronic, and untranslated regions creates an opportunity to increase our knowledge of genetic variation in noncoding gene regions and its relevance for inherited disorders.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:*

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: Swiss National Science Foundation grant no. 310000_119839/1 to C.R. Largiadèr and R. Jaggi.
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

Acknowledgments: The authors thank Marcelo Caraballo, John Cooper, and Sandra Munro for helpful discussions on the MSC procedure, as well as Jean-Marc Nuoffer and 2 anonymous reviewers, for their constructive comments on the manuscript. Special thanks to Karim Gharbi and the GenePool at the University of Edinburgh, as well as to Christof Wunderlin, for their support concerning 454 sequence data analysis.

References

1. Mian A, Lee B. Urea-cycle disorders as a paradigm for inborn errors of hepatocyte metabolism. Trends Mol Med 2002;8:583–9.
2. Endo F, Matsuura T, Yanagita K, Matsuda I. Clinical manifestations of inborn errors of the urea cycle and related metabolic disorders during childhood. J Nutr 2004;134:1605S–9S.
3. Yamaguchi S, Brailey LL, Morizono H, Bale AE, Tuchman M. Mutations and polymorphisms in the human ornithine transcarbamylase (OTC) gene. Hum Mutat 2006;27:626–32.
4. Engel K, Nuoffer J-M, Mühlhausen C, Klaus V, Largiadèr CR, Tsiekas K, et al. Analysis of mRNA transcripts improves the success rate of molecular genetic testing in OTC deficiency. Mol Genet Metab 2008;94:292–7.
5. Ogino W, Takeshima Y, Nishiyama A, Akizuka Y, Yagi M, Tsuneishi S, et al. Mutation analysis of the ornithine transcarbamylase (OTC) gene in five Japanese OTC deficiency patients revealed two known and three novel mutations including a deep intronic mutation. Kobe J Med Sci 2007;53:229–40.
6. Azevedo L, Soares PA, Quental R, Vilarinho L, Teles EL, Martins E, et al. Mutational spectrum and linkage disequilibrium patterns at the ornithine transcarbamylase gene (OTC). Ann Hum Genet 2006;70:797–801.
7. Mak CM, Siu T-SS, Lam C-WW, Chan GC, Poon GW, Wong K-YY, et al. Complete recovery from acute encephalopathy of late-onset ornithine transcarbamylase deficiency in a 3-year-old boy. J Inherit Metab Dis 2007;30:981.
8. Scaglia F, Zheng Q, O'Brien WE, Henry J, Rosenberger J, Reeds P, Lee B. An integrated approach to the diagnosis and prospective management of partial ornithine transcarbamylase deficiency. Pediatrics 2002;109:150–2.
9. Häberle J, Denecke J, Schmidt E, Koch HG. Diagnosis of *N*-acetylglutamate synthase deficiency by use of cultured fibroblasts and avoidance of nonsense-mediated mRNA decay. J Inherit Metab Dis 2003;26:601–5.
10. Häberle J, Koch HG. Genetic approach to prenatal diagnosis in urea cycle defects. Prenat Diagn 2004;24:378–83.
11. Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res 2008;18:1638–42.
12. Chou L-S, Liu C-SJ, Boese B, Zhang X, Mao R. DNA sequence capture and enrichment by microarray followed by next-generation sequencing for targeted resequencing: neurofibromatosis type 1 gene as a model. Clin Chem 2010;56:62–72.
13. ten Bosch JR, Grody WW. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. J Mol Diagn 2008;10:484–92.
14. Raca G, Jackson C, Warman B, Bair T, Schimmenti LA. Next generation sequencing in research and diagnostics of ocular birth defects. Mol Genet Metab 2010;100:184–92.
15. Vasta V, Ng SB, Turner EH, Shendure J, Hahn SH. Next generation sequence analysis for mitochondrial disorders. Genome Med 2009;1:100.
16. Hoischen A, Gilissen C, Arts P, Wieskamp N, van der Vliet W, Vermeer S, et al. Massively parallel sequencing of ataxia genes after array-based enrichment. Hum Mutat 2010;31:494–9.
17. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007;4:903–5.
18. Bau S, Schracke N, Kränzle M, Wu H, Stähler PF, Hoheisel JD, et al. Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. Anal Bioanal Chem 2009;393:171–5.
19. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. Nat Methods 2007;4:907–9.
20. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009;27:183–9.
21. Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M. Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res 2007;35:e97.
22. D'Ascenzo M, Meacham C, Kitzman J, Middle C, Knight J, Winer R, et al. Mutation discovery in the mouse using genetically guided array capture and resequencing. Mamm Genome 2009;20:424–36.
23. Strömberg M, Lee W-P. MOSAIK [computer software]. Version 1.0.1370. http://bioinformatics.bc.edu/marthlab/Mosaik (Accessed January 2010).
24. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2009.
25. Sherry S, Ward M, Kholodov M, Baker J, Phan L, Smigielski E, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–11.
26. Biesecker LG. Exome sequencing makes medical genomics a reality. Nat Genet 2010;42:13–4.
27. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science 2009;323:133–8.
28. Summerer D, Wu H, Haase B, Cheng Y, Schracke N, Stähler CF, et al. Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. Genome Res 2009;19:1616–21.
29. Hoppman-Chaney N, Peterson LM, Klee EW, Middha S, Courteau LK, Ferber MJ. Evaluation of oligonucleotide sequence capture arrays and comparison of next-generation sequencing platforms for use in molecular diagnostics. Clin Chem 2010;56:1297–306.
30. Carletti E, Guerra E, Alberti S. The forgotten variables of DNA array hybridization. Trends Biotechnol 2006;24:443–8.
31. Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, et al. Human catechol-*O*-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science 2006;314:1930–3.
32. Lee H, O'Connor BD, Merriman B, Funari VA, Homer N, Chen Z, et al. Improving the efficiency of genomic loci capture using oligonucleotide arrays for high throughput resequencing. BMC Genomics 2009;10:646.